گزارش کار پروژه دوم درس داده‌کاوی محاسباتی

استخراج کلمات و جملات کلیدی

محمد لشکری ۴۰۰۱۱۲۰۸۷

دادگان

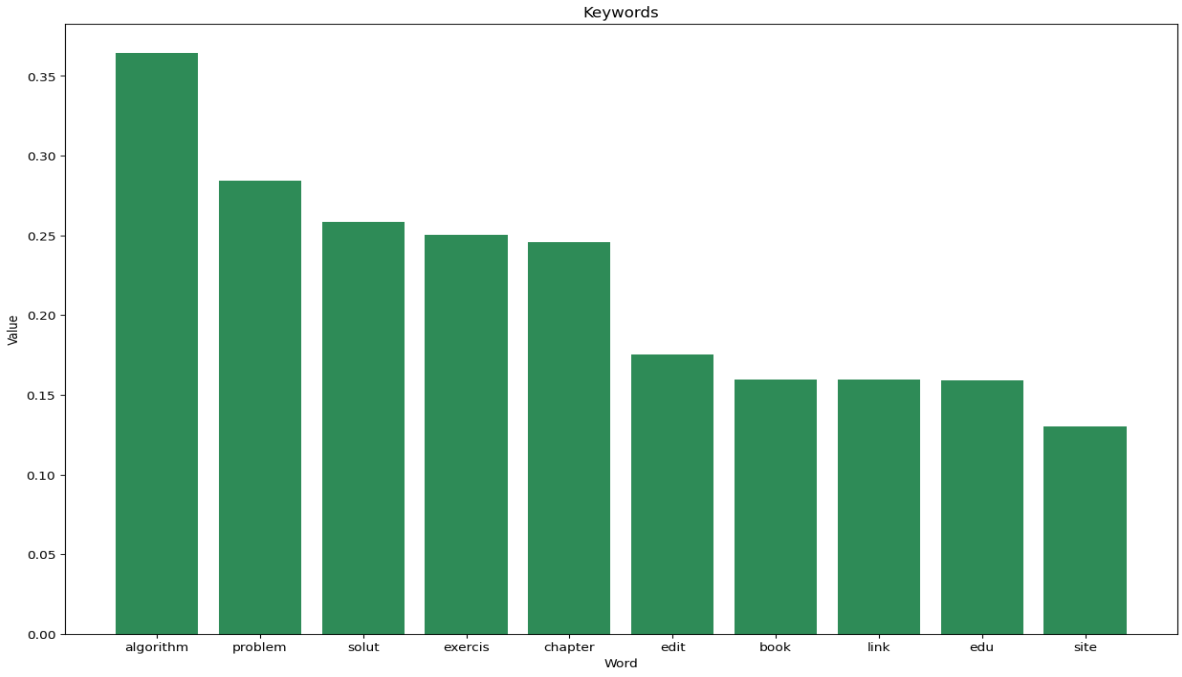
بخش preface کتاب introduction to algorithms(3rd edition) که به CLRS معروف است، به عنوان متن ورودی در نظر گرفته شد. این بخش از این کتاب، حدود هفت صفحه و شامل ۱۵۶ جمله است.

پیش پردازش دادگان

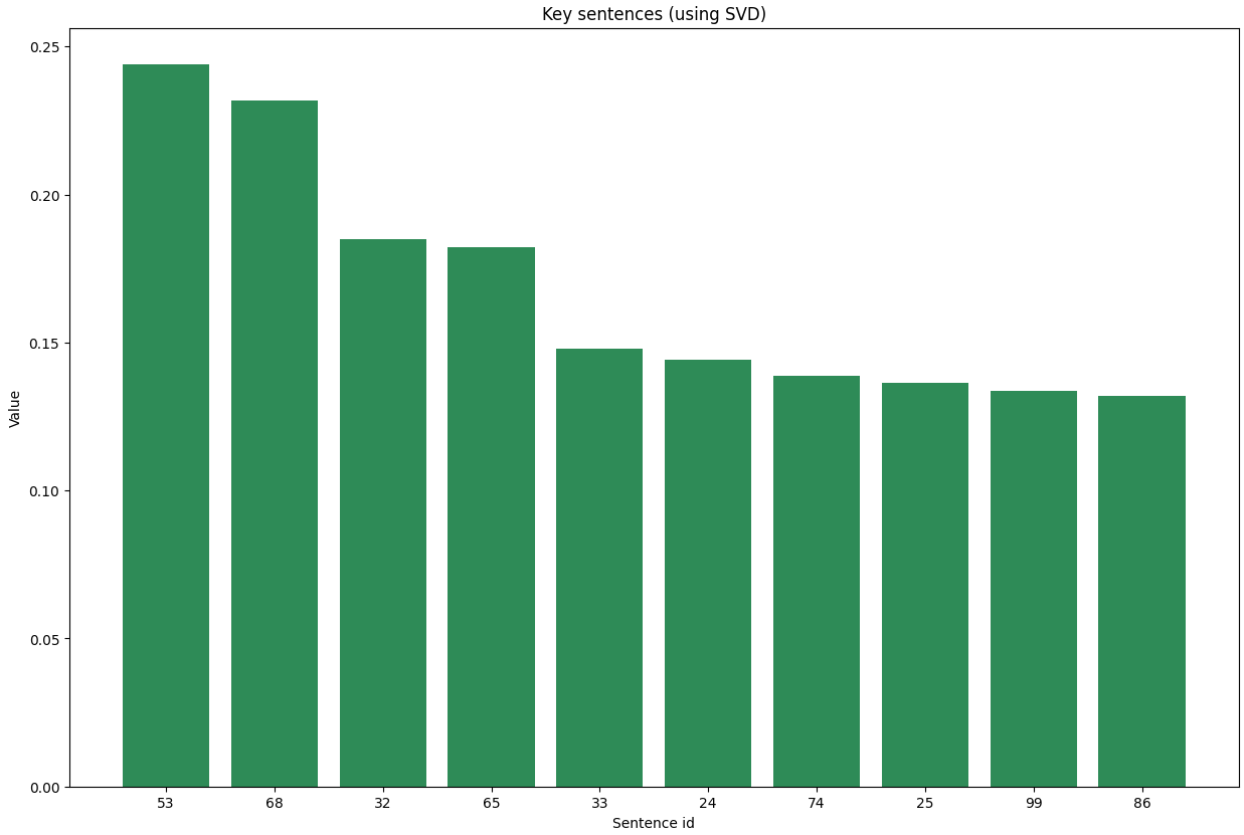
تمامی کاراکترهای اضافی به جز حروف انگلیسی، اعداد، علامت سؤال و علامت تعجب از دادگان حذف شدند. سپس کلمات بسیار پرتکرار مانند a و an و the و of از دادگان حذف شدند و درنهایت، کلمات به ریشه خود تبدیل شدند[1]؛ زیرا بیانگر یک ویژگی هستند.

روش ساده برای استخراج کلمات و جملات (استفاده از تجزیه مقدار منفرد)

فرض کنیم A ماتریس هم‌وقوعی باشد. با استفاده از تجزیه مقدار منفرد، A به صورت $A = U\Sigma V^T$ تجزیه می‌شود. ستون اول ماتریس U لغات و ستون اول V جملات را به ترتیب اهمیت مرتب می‌کند. ۱۰ لغت و ۱۰ جمله برتر در نمودارهای زیر قابل مشاهده است:



[1] Word stemming (e.g. the word “computing” has been converted to “compute”.)



سه جمله برتر بدست آمده توسط این روش به صورت زیر هستند:

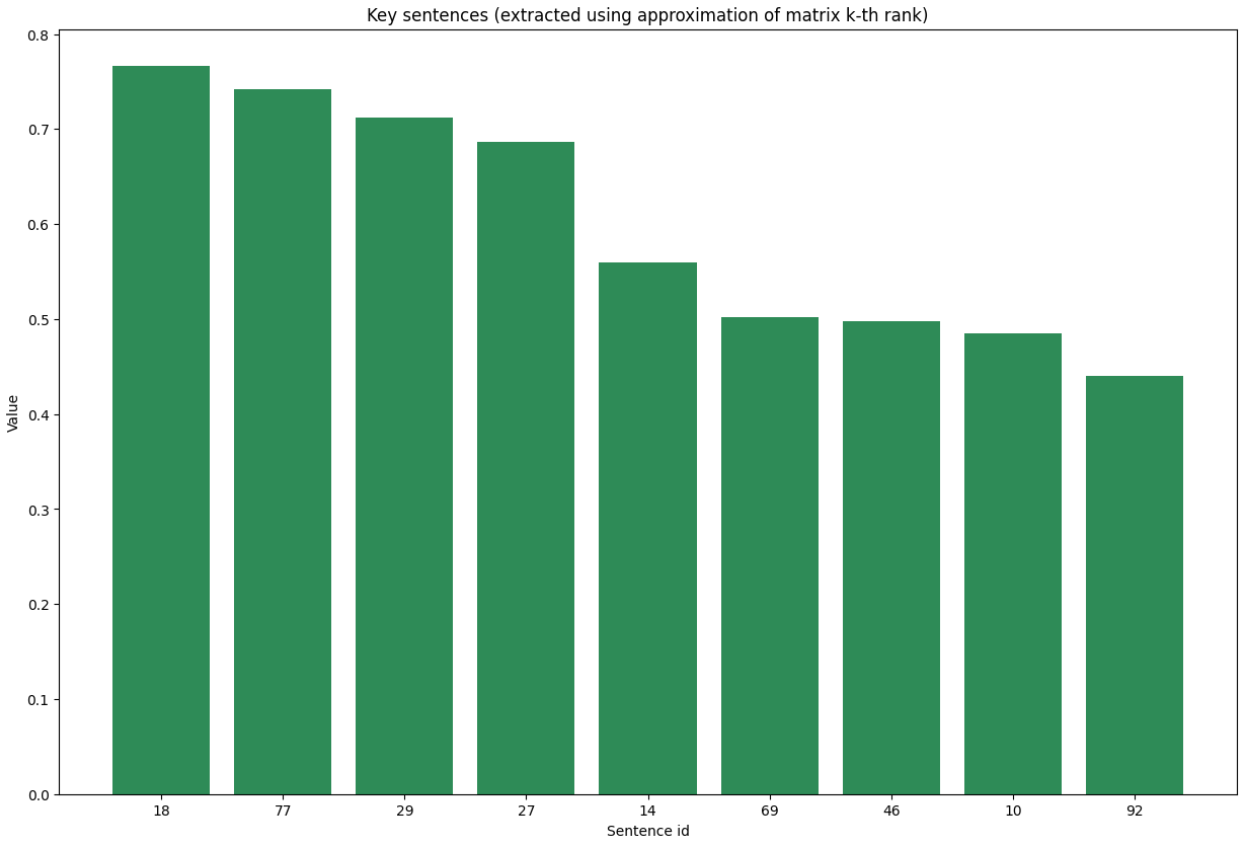
```
edu/algorithms/, links to solutions for a few of the problems and exercises.
====================
edu/algorithms/, links to solutions for some of the problems and exercises so
that you can check your work.
 ====================
edu/algorithms/, links to these solutions.
```

همان‌طور که انتظار می‌رفت، در نتیجه حاصل از این روش جملات با کلمات مشابه ظاهر شدند. در ادامه با پیاده‌سازی روش تقریب رتبه k-ام ماتریس هم‌وقوعی سعی می‌کنیم این مشکل را برطرف کنیم.

استخراج جملات کلیدی با روش تقریب رتبه k-ام ماتریس

همان‌طور که در بالا مشاهده می‌شود، جملات به دست آمده از صفحه قبل مشابه یکدیگراند و یک مفهوم را می‌رسانند. برای رفع این مشکل می‌توان به جای تجزیه مقدار منفرد از تقریب رتبه k-ام ماتریس استفاده کرد که با تجزیه نامنفی ماتریس (با k مؤلفه) آغاز می‌شود و در حداکثر k مرحله، حداکثر k جمله کلیدی را برمی‌گرداند.

نتایج این روش برای k=10 در نمودار زیر آمده است:



سه جمله برتر بدست آمده توسط این روش به صورت زیر هستند:

```
This is a large book, and your class will probably cover only a portion of its
material.

====================

A quick look at the table of contents shows that most of the second-edition
chapters and sections appear in the third edition.

====================

Departing from our practice in previous editions of this book, we have made
publicly available solutions to some, but by no means all, of the problems and
exercises.
```

همان‌طور که مشاهده می‌شود، مجموعه جملات فوق خلاصه بهتری برای متن اصلی ارائه می‌دهند. زیرا در این روش، بعد از انتخاب یک جمله، وزن جملات مشابه آن به نحوی تغییر می‌کند(کوچک می‌شود) که دیگر درمرحله بعد انتخاب نشوند.